基于 LDA 模型的移动投诉文本热点话题识别*

方小飞[1]　黄孝喜[1]　王荣波[1]　谌志群[1]　王小华[1,2]
[1](杭州电子科技大学计算机学院　杭州　310018)
[2](中国计量大学　杭州　310018)

摘要:【目的】运用中文信息处理和话题识别与追踪的方法，从大量移动投诉文本中找出有价值的信息。【方法】从分析投诉文本的特点入手，使用 k-means 先对文本聚类。利用 LDA 对每个类进行建模，提取话题，并从词频、词跨度和词长三方面计算每个话题中词的权值，把权重最大的词作为该话题的标签，并计算每个话题的文档分布概率均值。对具有相同标签的话题，先按照均值最大的原则去掉重复标签话题，再对所有话题计算文档支持率，并将文档支持率作为话题的热度，通过热度区分热点话题和一般话题。【结果】对投诉文本进行时间上的建模，通过对比一般话题和热点话题，得出热点话题的支持文档率至少是一般话题的 3 倍，支持文档率变化趋势也比一般话题高，说明本文算法是有效的。【局限】没有考虑到话题之间的语义关系。【结论】利用 LDA 模型对移动投诉话题检测初探的方法是比较合理和有效的，对今后此领域的研究具有一定的借鉴意义。

关键词: 移动投诉　k-means　话题识别　LDA 模型
分类号: TP391

1 引　言

随着互联网的日益普及和通信技术的不断发展，上网人群日益增多，尤其是移动端，通过手机玩游戏、刷微博、逛贴吧、看新闻的人越来越多。各大电信运营商看到了其中的商机，为了满足客户的需求、拓宽自己的业务、抢占市场份额，他们推出了各种优惠政策以吸引用户，随着用户量不断上升，投诉量也日益剧增，因此如何有效地处理投诉文本成为了各界关注的焦点。其实在大量的投诉中有很多大家关注的热点话题，比如“宽带”、“流量”、“扣费”等等，如果可以从中发现话题，并对话题进行追踪，根据话题的变化趋势了解相关业务的受理情况、了解用户的关注点，从而对症下药，就能提高处理投诉的效率。因此对投诉文本进行话题挖掘就显得十分重要。

与新闻报道相比较，移动投诉文本的结构更加复杂且短小，这加大了提取话题的难度。本文针对移动投诉文本，应用“话题识别”的相关知识，从中识别投诉文本中的热点话题。

2 相关工作

话题识别和跟踪研究中，LDA[1]主题模型是近年来文本挖掘领域的一个热门研究方向，主题模型具有优秀的降维能力、针对复杂系统的建模能力和良好的扩展性。利用主题建模挖掘出的主题可以帮助人们理解海量文本背后隐藏的语义，也可以作为其他文本分析方法的输入，完成文本分类、话题检测、文本自动摘要和关联判断等多方面的文本挖掘任务。

通讯作者: 黄孝喜，ORCID: 0000-0003-4483-3664，E-mail: huangxx@hdu.edu.cn。

*本文系国家自然科学基金青年基金项目“引入涉身认知机制的汉语隐喻计算模型及其实现”(项目编号: 61103101)、国家自然科学基金青年基金项目“基于马尔科夫树与 DRT 的汉语句群自动划分算法研究”(项目编号: 61202281)和教育部人文社会科学研究青年基金项目“面向信息处理的汉语隐喻计算研究”(项目编号: 10YJCZH052)的研究成果之一。

LDA 主题模型具有优秀的降维能力和扎实的概率理论基础，使其在短文本主题挖掘中具有很大的潜力。近年来，为了提高 LDA 模型主题挖掘的效率和准确性，出现很多对 LDA 模型的改进方法，可归纳为纵向的过程扩展和横向的模型扩展[2]。一方面，针对微博文本篇幅较短的局限，基于操作过程扩展的方法考虑将微博文本进行适当的聚集，这样短文本被聚集成相对适合挖掘的长文本。Weng 等[3]采用同一微博用户的所有微博文本聚集成一篇长文档的策略，利用LDA模型进行主题挖掘。Hong 等[4]提出基于训练的用户模式建模和基于术语模式建模。另一方面，为了适应微博短文本的主题挖掘，规避短文本数据噪声大的影响，提出基于模型扩展优化的 LDA 模型，典型的改进模型包括 ATM[5]、Twitter-LDA[6]、Labeled-LDA[7]、MB-LDA[8]、HLDA[9]以及 MA-LDA[10]。Zhao 等[6]提出 Twitter-LDA 模型挖掘整个 Twitter 文本中具有代表性的文本主题。Ramage 等[7]提出 Labeled-LDA，一种基于标签的主题模型。张晨逸等[8]提出微博生成模型 MB-LDA，该模型综合考虑了微博的文本关联关系和联系人关联关系，这两种关系可以辅助微博的主题挖掘。文献[2, 11]对 LDA 模型的纵向和横向改进方法进行了比较总结，如表 1 所示。

表 1　LDA 话题模型建模方法比较[2]

模型	扩张方式	实现方式	优势	局限性
LDA[1]	无	直接使用	无需监督	主题挖掘不理想
基于用户聚集 LDA[3]	过程扩展	文本聚集	解决短文本问题	只限微博用户层面建模，需要人工干预
基于训练 USER 模式[4]	过程扩展	文本聚集、分步求解	解决短文本问题，简化推导	需要事先训练和人工干预，若要更新模型需重新训练基
ATM[5]	模型扩展	文本聚集	解决短文本问题	只限微博用户层面主题建模
ATM 扩展模型[12]	模型扩展	文本聚集	解决短文本问题	帖子层面主题少且不理想
Twitter-LDA[6, 13]	模型扩展	文本聚集，引入背景模型	解决短文本问题和高频词汇问题	一个帖子只能对应一个主题
Labeled-LDA[7, 14]	模型扩展	引入标签信息	提高主题可解释性	要求文本具有足够的标签信息
MB-LDA[8]	模型扩展	引入结构化信息	解决短文本问题，提高主题可解释性	主要针对会话类和转发类中文微博
HLDA[9]	模型扩展	引入微博评论数、转发数等特征量	提高主题可解释性	主要针对具有高评论数和转发数的微博
MA-LDA[10]	模型扩展	引入时间特征	解决短文本问题，提高主题可解释性	主要适应于短时间内被普遍关注的微博

本文鉴于 LDA 模型本身的优点和在短文话题识别上的优势，又考虑到投诉文本与微博短文本不一样，微博一般围绕一个话题展开，包含评论、转发等额外信息；但投诉文本没有一个明确的话题，仅仅是客户的一条信息反馈，文本结构简短，内容复杂。因此，本文提出一种基于 LDA 模型的移动投诉文本热点话题识别方法。首先对投诉文本聚类，每一类使用 Gibbs 抽样方法进行话题的抽取；然后对抽取的话题进行一系列的处理；最后通过计算话题的文档支持率得出热点话题，并在实验部分对本文方法进行了验证。

3　基于 LDA 模型的移动投诉文本热点话题识别

3.1　文本聚类

由于投诉文本跟新闻报道不一样，它的形式简短，单条文本涵盖内容信息很少。为了更好的提取话题，首先将文本进行聚类，这样每一类中的投诉文本不仅存在着共性，而且内容比较充实，LDA 模型抽取话题表达效果就会更好，针对性更强。

本文采用 k-means[15]进行聚类，k-means 是经典划分聚类算法。这种方法简单快速，在对文档进行聚类

前需要通过 k 值来确定簇数量。主要过程是从含 n 个文本的文档集中随机选择 k 个文本作为初始的聚类中心，并通过计算得到其他文本到每个簇中心点的距离，将文档划分到离它最近的簇中，用迭代的方式不断重复上述过程，直到满足准则函数或划分过程中相邻簇的中心不再发生变化为止。通过不断的迭代过程增加簇内的紧凑性，降低簇间的相似性。图 1 为本文聚类的流程。

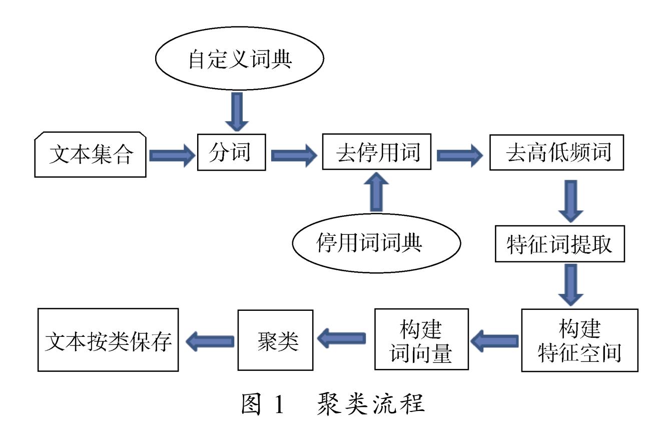


图 1　聚类流程

使用 k-means 聚类好后，将投诉文本按各类分别存于一个 Txt 文件中。

3.2　LDA 模型话题抽取

LDA[1]模型中对话题的定义为：一组语义上相关的词及这些词在该话题上的分布概率。由于无法对 LDA 模型的未知参数进行求解，在这里使用 Gibbs Sampling 的方法近似求解，Gibbs Sampling[16]通过迭代采样达到逼近真实结果的效果，其关键点在于对当前单词采样概率的求解，如公式(1)[1]所示。

$$P(z_i^d)\,|\,z_{-i}^d,w)=\infty\frac{C_{dj}^{DK}+\alpha}{\prod_{K=1}^{K}C_{dk}^{DK}+K\alpha}\frac{C_{ij}^{VK}+\beta}{\prod_{K=1}^{K}C_{ki}^{VK}+V\beta}\qquad(1)$$

其中，w 为词表个数；K 为话题数目；C_{ij}^{VK} 为计数矩阵 C^{VK} 中第 ij 项，表示第 j 个话题中第 i 个词出现的次数；C_{dj}^{DK} 为计数矩阵 C^{DK} 中的第 dj 项，表示第 d 篇文档中，第 j 个话题包含的词的数目。通过 Gibbs Sampling 方法，可以得到 θ、ϕ 的后验值，如公式(2)[1]和公式(3)[1]所示。

$$\theta_j^d=\frac{C_{dj}^{DK}+\alpha}{\prod_{k=1}^{k}C_{dk}^{DK}+K\alpha}\qquad(2)$$

$$\varphi_j^i=\frac{c_{ij}^{VK}+\beta}{\prod_{k=1}^{v}c_{kj}^{VK}+v\beta}\qquad(3)$$

在推导参数之前，需要预先将话题的数目 K 设置好，数值越大则话题越多，话题的颗粒度越小，反之亦然。K 的取值对 LDA 模型文本提取和拟合性能影响较大，其最佳的确定可以通过两种方法：一种是词汇被选中的概率 $p(w|T)$[17]，另一种是困惑度(perplexity)[17]。本文用困惑度确定 K，困惑度越小，话题的拟合性就越好。困惑度计算如公式(4)[17]所示。

$$perlexity=\exp\left(-\frac{\sum_{i=1}^{M}\log(p(d_i))}{\sum_{i=1}^{M}N_i}\right)\qquad(4)$$

其中，M 为文本数，N_i 为文本 d_i 的长度(即单词个数)，$p(d_i)$ 为 LDA 模型产生文本 d_i 的概率。

3.3　热点话题识别

使用 Gibbs Sampling 抽样可以得到"话题–词语"和"文档–话题"的概率分布。对于"话题–词语"分布，每个话题 z 下分布着词语 w 和它在此话题中的概率 $p(w|z)$，话题 $z=\{(w_1,p(w_1\,|\,z)),\cdots,(w_i,p(w_i\,|\,z)),\cdots,(w_n,p(w_n\,|\,z))\}$ 对于"文档–话题"分布，每个文档 d 下分布着 k 个话题的概率分布，形如 $D=\{P(z_1\,|\,d),\cdots P(z_i\,|\,d)\cdots,P(z_k\,|\,d)\}$。

使用 Gibbs Sampling 抽取的话题数量会比较多，而且有些话题可能表达的意思十分接近，有些话题几乎不能表达文档的意思，所以要进行话题选取。话题的选取就要用到上面的"话题–词语"和"文档–话题"的概率分布。经过话题选取之后，确定了文本的全局话题，然后从全局话题中发现热点话题。

(1) 选取话题标签词

文本经过聚类，得到了 H 个类，每个类使用 Gibbs Sampling 得到了若干个隐含的话题，每个话题下分布着 n 个话题相关的词，对每个话题中的词计算其在该话题所在类文本中的词频(*count*)、词跨度(*cover*)和词的长度(*length*)，则该词的权值(*weight*)计算公式如(5)所示。

$$weight=count+length+cover\qquad(5)$$

为了不让词频、词的长度和词跨度的值相差太大，使三者在权值中的比重相同，分别对其进行了量化，具体计算如公式(6)–公式(8)所示。

$$count=\frac{count(i)}{count(i)+1} \tag{6}$$

$$length=\frac{length(i)}{\max(length(j))} \tag{7}$$

$$cover=\frac{last(i)-first(i)}{ctotal} \tag{8}$$

其中，*count*(*i*)为词在文档出现的次数，*length*(*i*)为词的长度，max(*length*(*i*))为文档中词的最大长度，*last*(*i*)为词在文档中最后一次出现的位置，*first*(*i*)为词在文档中第一次出现的位置，*ctotal* 是文档中最后一个词的位置。计算完话题中词的权值后，选出权值最大的词作为该话题的标签词。然后存入数据库，数据表的字段名为标签词(*tag*)、话题(*topic*)和话题所表示的类(*H*)。

(2) 计算话题的文档概率分布均值

通过 Gibbs Sampling 对每个类抽样后，各自得到一个“文档–话题”概率分布矩阵，矩阵表达式如公式(9)所示。

$$\begin{array}{c|ccccc|} & z_1 & \cdots & z_i & \cdots & z_k \\ d_1 & p(z_1|d_1) & \cdots & p(z_i|d_1) & \cdots & p(z_k|d_1) \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ d_i & p(z_1|d_i) & \cdots & p(z_i|d_i) & \cdots & p(z_k|d_i) \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ d_m & p(z_1|d_m) & \cdots & p(z_i|d_m) & \cdots & p(z_k|d_m) \end{array} \tag{9}$$

上述矩阵中有 *k* 个话题和 *m* 条文档，每行为 *k* 个话题在一条文档中的分布概率，每列为一个话题在 *m* 个文档中的分布概率。通过上面的矩阵概率分布就可以得出每个话题的分布概率均值，具体计算如公式(10)所示。

$$\mathrm{AVG}(Z_i)=\frac{\sum_{j=1}^{m}p(z_i|d_j)}{m} \tag{10}$$

(3) 话题选取

得到了话题的标签词和话题的文档概率分布均值后，构建话题矩阵如公式(11)所示。

$$\begin{array}{c} topic_1: \\ \vdots \\ topic_i: \\ \vdots \\ topic_n: \end{array}\begin{bmatrix} topic_1_tag & \cdots & avg(topic_1) & \cdots & H_I \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ topic_i_tag & \cdots & avg(topic_i) & \cdots & H_J \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ topic_n_tag & \cdots & avg(topic_n) & \cdots & H_K \end{bmatrix} \tag{11}$$

矩阵中一共有 *n* 个话题(topic)，$topic_i_tag$ 为 $topic_i$ 的标签词，avg($topic_i$)为 $topic_j$ 的文档概率分布均值，H_I、H_J 和 H_K 属于文本类集合 H。由于话题标签词存在相同的情况，所以先以话题标签词分组。认为同一组中的话题表达的意思相近，如果一组中有多个话题选取其中分布概率均值最大的话题，将其删除。接下来按每个话题的均值排序，去除均值极小的话题，因为均值小的话题不能很好地表达文档的意思，剩下的话题就是文档的全局话题。

(4) 热点话题识别

根据 LDA 模型的原理，每篇文档都是由数个不同的话题按照一定的比例生成的。这里假设一条经过预处理的投诉文本中有不少于话题 *z* 中百分之几的词，则认为这条投诉文本是话题 *z* 的支持文档。之后使用徐佳俊等[18]的方法计算文档话题支持率，如公式(12)[18]所示。如果在一个时间段内，话题的支持文档的数量或者文档话题支持率超过一个设定的阈值，那么这个话题就是热点话题。

$$S(z,t)=\frac{|D_i^t|}{|D^t|} \tag{12}$$

其中，*z* 表示话题，*t* 表示时间段，$|D_i^t|$ 为时间段 *t* 内话题 *z* 的所有支持文档数，$|D^t|$ 为时间段 *t* 内所有文档数量。

通过箱型图分析[21]进行话题支持文档数或者文档支持率阈值的设定，箱型图的结构如图 2 所示。

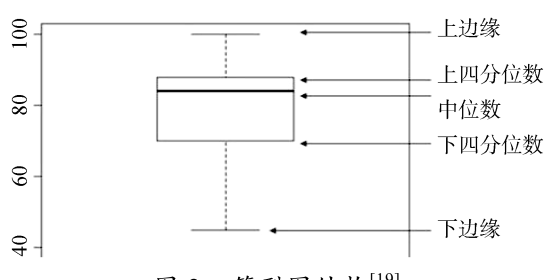


图 2　箱型图结构[19]

箱型图用来分析数据的分布情况和识别异常值。从图 2 中可以看出数据分为 4 个部分，位于上边缘之上和下边缘之下的值为异常值，本文不作考虑。这里将上四分位数这个值设定为支持文档数的阈值，如果某个话题的支持文档数的值超过这个阈值，该话题为热点话题。一方面，箱形图的绘制依靠实际数据，不需要事先假定数据服从特定的分布形式，没有对数据作

任何限制性要求，它只是真实直观地表现数据形状的本来面貌；另一方面，箱形图判断异常值的标准以四分位数和四分位距为基础，四分位数具有一定的耐抗性，多达 25%的数据可以变得任意远而不会很大地扰动四分位数，所以异常值不能对这个标准施加影响，箱形图识别异常值的结果比较客观。

4 实验及结果分析

4.1 数据来源

本文所使用的数据是某电信公司投诉业务部提供的，实验部分使用 2015 年 3 月–2015 年 4 月的投诉文本，其中，3 月份有 20 000 多条，4 月份有 50 000 多条，前者用于训练提取话题和识别热点话题，后者用于验证热点话题抽取的效果。分词使用的是结巴分词工具[20]，停用词词典为哈尔滨工业大学的停用词词典[21]。

4.2 语料预处理

(1) 由于现有的词典无法完全识别投诉业务中的专业术语和业务词，为了提高分词效果，在某电信公司业务部员工的协助下手动建立了一个自定义的分词词典，词典一共包含了 1 600 个重点业务关键词，由三元组(词语，词频，词性)组成，其中词性标注集采用的是中国科学院的汉语文本词性标注集。三元组中各个属性以空格分开，每个三元组独占一行，保存在 Txt 文件中。词典实例如表 2 所示。

表 2 词典实例表

词语	词频	词性
短信费用	1 000	n(名词)
欠费停机	2 000	n
上网费用	2 000	n
有线宽带	2 000	n
畅玩游戏包	500	n
爱动漫信息费	3 000	n
夜间流量	28 641	n

(2) 使用正则表达式去除投诉文本中特有的短语，例如“手机号码”、“工单号”等由字母和数字组成的字符串。

(3) 引入自定义词典，使用结巴分词工具进行分词并标注词性，保留名词、动词等重要的词语，并去除停用词。

(4) 去除无关的高频词，由于投诉文本是由专业的服务人员使用软件按照模板格式录入的，所以会有很多无法反映语义信息的重复词，例如“诉求”、“用户来电表示”、“客户资产编号”、“请处理”、“谢谢”等，预处理的流程如图 3 所示。

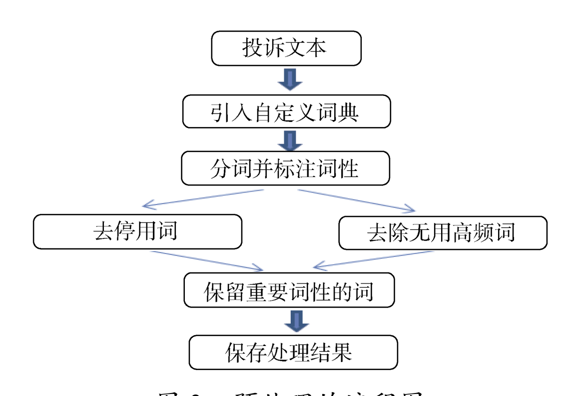


图 3 预处理的流程图

处理结果示例如表 3 所示。

表 3 语料预处理效果表

原始文本	诉求：用户来电表示自己的手机(18067938538)自己很少上网，为什么在(2015-03)月份的手机上会超出(210.38)兆的上网流量，前台解释不使用是不会产生这个上网流量，前台建议用户手机不使用的时候把数据流量的开关关闭，用户可以登录网厅查询上网的详单，前台解释用户不认可，用户表示自己没有使用产生的流量费用自己也是不予承担，烦请处理谢谢，客户资产编号：1-14PTKKXU
处理结果	手机 上网 月份 手机 超出 上网流量 前台 解释 上网流量 前台 建议 手机 数据流量 开关 关闭 登录 网厅 查询 上网 详单 前台 解释 用户不认可 流量 费用 不予 承担

4.3 聚 类

通过采用模糊 k-means 聚类，k 设置为 200，并对聚类结果中每类的文本条数进行了统计，其中条数最少为 45 条，最多的有 362 条。具体如表 4 所示。

表 4 文本类分布表

类中文本条数区间	类个数
[0, 50]	20
[51, 100]	62
[101, 200]	88
[201, +∞)	40

4.4 全局话题抽取

实验利用 Gibbs Sampling 方法进行参数推理，使用基于 Java 的 Gibbs Sampling 开源工具包(JGibbLDA-v.1.0)[22]，模型参数 α、β 默认值为 50/k 和 0.1，每个话

题下的词语个数设置为 10。

对于话题个数 k，这里使用公式(4)进行计算，并对生成的话题进行人工评判。根据每类中文本的条数，最终认定条数在[0, 50]区间内的类，k 值设置为 5；条数在[51, 100]区间内的类，k 值设置为 10；条数在[101, 200]区间内的类，k 值设置为 20；条数在[201, 300]区间内的类，k 值设置为 30；条数在[301, +∞)区间内的类，k 值设置为 40。

参数 α、β 和 k 设置好后，对每类进行话题的抽取，得到”话题–词语”和“文档–话题”的概率分布，如表 5 和图 4 所示。

表 5　话题–词语示例表

话题	词语及词概率
Topic 0th:	金额 0.494　号码 0.056　宽带 0.042　接到 0.028　核实 0.028　收费 0.015　订单 0.015　114 0.015　显示 0.015　退订 0.015
Topic 1th:	违约金 0.228　不认可 0.122　无 0.046　对此 0.046　滞纳金 0.031　卡里 0.031　用户不认可 0.031　费用 0.031　通知 0.031　收费 0.016
Topic 2th:	违约金 0.092　返利 0.077　翼支付 0.062　投诉 0.062　成功 0.031　翼支付加油 0.031　支付 0.031　收到 0.031　营业 0.031　办理 0.031
Topic 3th:	滞纳金 0.33　欠费 0.101　电话 0.044　上门 0.030　违约金 0.030　加油 0.030　平台 0.030　前台 0.015　怎么回事 0.015　出票 0.015
Topic 4th:	减免 0.248　前台 0.087　解释 0.087　交易 0.062　强烈要求 0.038　账户 0.025　情况 0.025　营业厅 0.025　无效 0.013　号用 0.013

1	0.2222222222222222	0.2222222222222222	0.1527777777777778	0.2222222222222222	0.1805555555555555
2	0.16923076923076924	0.23076923076923078	0.26153846153846155	0.16923076923076924	0.16923076923076924
3	0.2037037037037037	0.2037037037037037	0.18518518518518517	0.2037037037037037	0.2037037037037037
4	0.1864406779661017	0.1694915254237288	0.288135593220339	0.1694915254237288	0.1864406779661017
5	0.19642857142857142	0.19642857142857142	0.17857142857142858	0.21428571428571427	0.21428571428571427
6	0.203125	0.1875	0.234375	0.1875	0.1875
7	0.21153846153846154	0.19230769230769232	0.19230769230769232	0.21153846153846154	0.19230769230769232
8	0.21311475409836064	0.21311475409836064	0.18032786885245902	0.19672131147540983	0.19672131147540983
9	0.19298245614035087	0.21052631578947367	0.17543859649122806	0.19298245614035087	0.22807017543859648
10	0.2	0.21666666666666667	0.18333333333333332	0.21666666666666667	0.18333333333333332
11	0.18556701030927836	0.1134020618556701	0.20618556701030927	0.20618556701030927	0.28865979381443296
12	0.19642857142857142	0.17857142857142858	0.19642857142857142	0.19642857142857142	0.23214285714285715
13	0.2	0.21666666666666667	0.2	0.18333333333333332	0.2
14	0.2	0.2	0.18181818181818182	0.21818181818181817	0.2
15	0.2037037037037037	0.2037037037037037	0.18518518518518517	0.2037037037037037	0.2037037037037037
16	0.22807017543859648	0.21052631578947367	0.19298245614035087	0.17543859649122806	0.19298245614035087
17	0.203125	0.203125	0.21875	0.203125	0.171875
18	0.18333333333333332	0.21666666666666667	0.16666666666666666	0.21666666666666667	0.21666666666666667
19	0.21153846153846154	0.19230769230769232	0.19230769230769232	0.21153846153846154	0.19230769230769232
20	0.16923076923076924	0.2153846153846154	0.18461538461538463	0.2	0.23076923076923078
21	0.19642857142857142	0.19642857142857142	0.17857142857142858	0.19642857142857142	0.23214285714285715
22	0.21818181818181817	0.18181818181818182	0.18181818181818182	0.2	0.21818181818181817
23	0.21153846153846154	0.19230769230769232	0.19230769230769232	0.21153846153846154	0.19230769230769232
24	0.1896551724137931	0.1724137931034483	0.1896551724137931	0.20689655172413793	0.2413793103448276
25	0.21153846153846154	0.19230769230769232	0.19230769230769232	0.21153846153846154	0.19230769230769232
26	0.2	0.2	0.21818181818181817	0.18181818181818182	0.2
27	0.1935483870967742	0.20967741935483872	0.1935483870967742	0.22580645161290322	0.1774193548387097
28	0.20689655172413793	0.20689655172413793	0.1724137931034483	0.1896551724137931	0.2241379310344828
29	0.23333333333333334	0.21666666666666667	0.2	0.16666666666666666	0.18333333333333332
30	0.21428571428571427	0.17857142857142858	0.19642857142857142	0.21428571428571427	0.19642857142857142
31	0.19402985074626866	0.208955223880597	0.22388059701492538	0.1791044776119403	0.19402985074626866

图 4　文档–话题示例

表 5 和图 4 是某个类的“话题–词语”和“文档–话题”的概率分布，表 4 中有每个话题的 10 个话题词及其词语的话题分布概率 $p(w|z)$，图 4 中每行为 5 个话题在一条文档中的分布概率，每列为一个话题在 31 条文档中的分布概率。

通过话题的选取，总共抽取了 5 130 个话题，然后对每个话题提取它的标签词，计算文档概率均值，去除均值极小的话题，保留相同标签词中均值最大的话题，剩下 299 个全局话题，示例结果如图 5 所示。

t_topic	topic	theta
套餐	宽带/工作人员/用户不认可/营业厅/工单/固话移机/优惠/套餐/享受/移机	0.20845
电脑	电脑/话费/手机/套餐/海信/核实/金蛋/费用/退费/退货	0.20784
违约金	违约金/返利/翼支付/投诉/成功/翼支付加油/支付/收到/营业/办理	0.20159
收取	前台/收取/办理/告知/师傅/外线/后台/移机/号码/	0.20141
平板	平板/办理/杭州/告知/核实/支付/强烈要求/拿到/发现/	0.20026
移机	移机/收取/核实/套餐/减免/投诉/费用不认可/业务/区域/无效	0.20024
分组	分组/公众/广电/频道/改回/节目/高清/机顶盒/后台/故障	0.20015
减免	减免/前台/解释/交易/强烈要求/账户/情况/营业厅/无效/号用	0.19944
金额	金额/号码/宽带/接到/核实/收费/订单//显示/退订	0.19836
发短信	发短信/手机/对此/退费/功能/收到/上网/联系电话/地址/更换	0.19820
邮箱	邮箱/号码/无效/功能/短信/账号/用户不认可/建议/亚洲/公司	0.19771
服务质量	服务质量/办理/对此/监督/答应/工单/导致/前期/游戏/免责	0.19765

图 5　话题示例

其中, t_topic 为话题标签, topic 为话题下的分布词语, theta 为话题在文档中分布概率均值。

4.5　热点话题识别实验结果分析

根据第 3 节的方法，假设一条经过预处理的投诉文本中有不少于话题 z 中一定比例的词，则认为这条投诉文本是话题 z 的支持文档。这里设置为 30%，因为实验中每个话题的词语个数为 10，即每条预处理后的投诉文本中词语与话题中词语的交集大于等于 3 个。通过分析话题支持文档数的箱型图如图 6 所示。得出结果如表 6 所示。

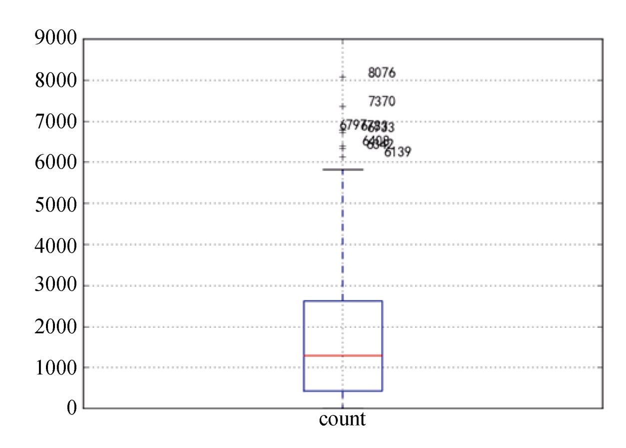


图 6　支持文档数箱型图

通过上述分析，这里将支持的文档数不低于 3 000 的话题定义为热点话题。由于话题个数有 299 个，文中分别选取了 10 个热点话题和 10 个一般话题进行说明，如表 7 所示。

表 6　支持文档数的数值分布

相关内容	数值
话题个数	299
支持文档数均值	1760.81
最大值	8076
最小值	3
中位数	1288
%25 分位数	424
%50 分位数	1288
%70 分位数	2628.5

表 7　话题对比表

热点话题	支持文档数	一般话题	支持文档数
账单	8 070	维修	1 643
用户不认可	6 797	包月费	1 058
副卡	6 733	违约金	526
短号	6 408	上门移机	428
路由器	6 342	服务质量	349
数据流量	5 360	一号双机	249
国内上网	4 848	翼支付	240
线路	4 262	租用	205
补卡	4 341	手机信号	55
宽带	3 225	彩铃	48

从表 7 中可以看出移动用户对"上网"、"数据流量"、"账单"等比较在意，这与现实中用户的关注基本符合，所以本文的话题抽取和热点识别方法是有效的。

4.6　话题测试实验结果分析

使用 2015 年 4 月的语料进行测试本文算法获取的热点话题效果，先按表 7 中话题支持文档数，从低至高分别选择三个热点话题和三个一般话题进行实验。计算热点话题和一般话题在 2015 年 4 月 30 天中的支持文档数，其变化趋势如图 7 和图 8 所示。

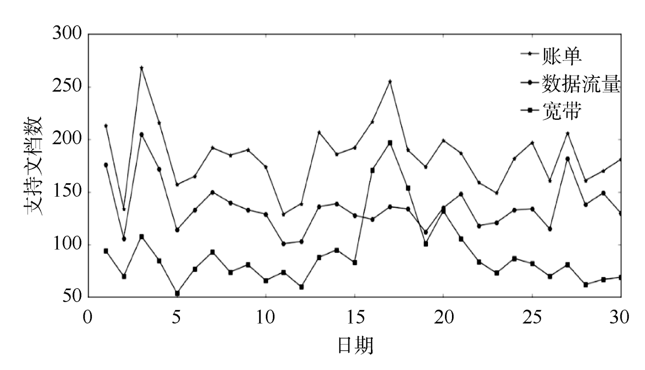


图 7　热点话支持文档数变化趋势

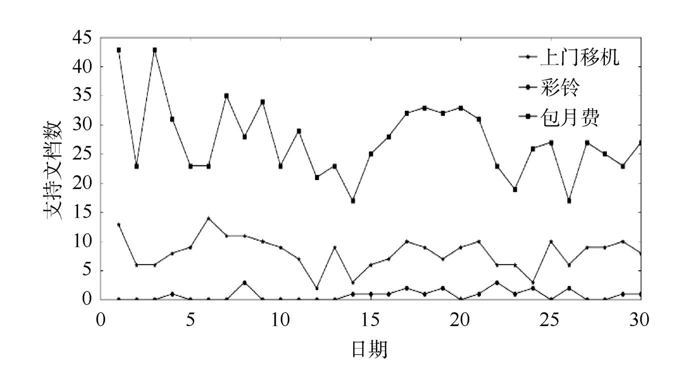


图 8　一般话题支持文档数变化趋势

对比图 7 和图 8 可以看出热点话题的每天支持文档数普遍比一般话题高；图 8 中一般话题的变化趋势大部分时间都比较平稳，有时也会出现急剧的爬升和下落，但支持文档数还是不高，最低点到最高点的变化幅度不是十分明显。从图 7 可以发现热点话题变化趋势强弱程度比较明显，最低点到最高点的变化幅度基本上都超过 100，有一个比较突出的峰值，总体都经历了"开始–高潮–衰落"的过程。

数据出现以上现象，从现实原因来说，是因为热点话题与用户的生活息息相关，都是大部分用户使用非常频繁的业务所出现的问题，所以它的强度变化趋势就比较明显。通过对不同话题进行趋势分析之后，可以发现它们的强度变化趋势与现实的实际情况是比较吻合的，在一定程度上能够反映本文算法获取热点话题的效果。

5　结　语

通过实验说明本文在基于投诉文本的热点话题识别问题研究中取得了一定成果。在预处理阶段，构建了一个移动领域的词典，对于今后该领域的语料处理有一定的帮助；在热点话题发现阶段，使用了聚类技术，使得类中的文本联系更加紧密；再通过 LDA 模型进行话题抽取，使话题表达更加细粒化，针对性更强；在话题的选取上，考虑了话题对文档表达能力的强弱以及话题与话题之间的相似性。

本文对移动投诉领域话题识别和追踪的初探，还存在一定的不足，没有考虑到话题之间的语义关系，使用的都是统计学的方法。接下将对此方法做出改善，把更多的语义信息融合到话题模型中；并对话题之间

的关系进行研究，发掘话题间的联系以及动态获取话题的演化。

参考文献：

[1] David M B, John D L.Dynamic Topic Model[C]//Proceedings of the 23rd International Conference on M achine Learning. Pittsburgh. 2006: 113-120.

[2] 张培晶，宋蕾. 基于 LDA 的微博文本主题建模方法研究述评[J]. 图书情报工作，2012，56(24)：120-126. (Zhang Peijing, Song Lei. Overview on Topic Modeling of Microblogs Text Based on LDA [J]. Library and Information Service, 2012, 56(24): 120-126.)

[3] Weng J, Lim E P, Jiang J, et al. TwitterRank: Finding Topic-sensitive Influential Twitterers[C]//Proceedings of the 3rd ACM International Conference on Web Search and Data Mining. ACM, 2010: 261-270.

[4] Hong L, Davison B D. Empirical Study of Topic Modeling in Twitter [C]//Proceedings of the 1st Workshop on Social Media Analytics. ACM, 2010: 80-88.

[5] Rosen-Zvi M, Griffiths T, Steyvers M, et al. The Author-Topic Model for Authors and Documents[C]// Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. AUAI Press, 2004: 487-494.

[6] Zhao W X, Jiang J, Weng J, et al. Comparing Twitter and Traditional Media Using Topic Models [C]// Proceedings of the 33rd European Conference on Information Retrieval. Springer Berlin Heidelberg, 2011: 338-349.

[7] Ramage D, Hall D, Nallapati R, et al. Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-labeled Corpora [C]//Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. 2009: 248-256.

[8] 张晨逸，孙建伶，丁轶群. 基于 MB-LDA 模型的微博主题挖掘[J]. 计算机研究与发展，2011，48(10)：1795-1802. (Zhang Chenyi, Sun Jianling, Ding Yiqun. Topic Mining for Microblog Based on MB-LDA Model [J]. Journal of Computer Research and Development, 2011, 48(10): 1795-1802.)

[9] 唐晓波，向坤. 基于 LDA 模型和微博热度的热点挖掘[J]. 图书情报工作，2014，58(5)：58-63. (Tang Xiaobo, Xiang Kun. Hotspot Mining Based on LDA Model and Microblog Heat [J]. Library and Information Service, 2014, 58(5): 58-63.)

[10] 朱颖. 基于微博的热点话题发现[D]. 重庆：西南大学，2014. (Zhu Ying. Hot Topic Extraction from Microblogs [D]. Chongqing: Southwest University, 2014.)

[11] 伍万坤，吴清烈，顾锦江. 基于 EM-LDA 综合模型的电商微博热点话题发现[J]. 现代图书情报技术，2015(11)：33-40. (Wu Wankun, Wu Qinglie, Gu Jinjiang. Hot Topic Extraction from E-commerce Microblog Based on EM-LDA Integrated Model [J]. New Technology of Library and Information, 2015(11): 33-40.)

[12] Rosen-Zvi M, Chemudugunta C, Griffiths T, et al. Learning Author-topic Models from Text Corpora [J]. ACM Transactions on Information Systems, 2010, 28(1): Article No.4.

[13] Zhao W X, Jiang J, He J, et al. Topical Key Phrase Extraction from Twitter [C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. 2011: 379-388.

[14] Ramage D, Dumais S T, Liebling D J. Characterizing Microblogs with Topic Models [C]//Proceedings of the 4th International Conference on Weblogs and Social Media. 2010.

[15] 吴夙慧，成颖，郑彦宁，等. K-means 算法研究综述[J]. 现代图书情报技术，2011(5)：28-35. (Wu Suhui, Cheng Ying, Zheng Yanning, et al. Survey on K-means Algorithm[J]. New Technology of Library and Information Service, 2011(5): 28-35.)

[16] 朱成文，李兵，胡奎. HMM 参数估计的 Gibbs 抽样算法[J]. 计算机工程与应用，2012，18(18)：57-60. (Zhu Chengwen, Li Bing, Hu Kui. Algorithm of Parameter Estimation of HMM via Gibbs Sampling. Computer Engineering and Applications, 2012, 48(18): 57-60.)

[17] 关鹏，王曰芬. 科技情报分析中 LDA 主题模型最优主题数的确定方法研究[J]. 现代图书情报技术，2016，32(9)：42-50.) (Guan Peng, Wang Yuefen. Identifying Optionan Topic Numbers from Sci-Tech Information with LDA Model[J]. New Technology of Library and Information, 2016, 32(9): 42-50.)

[18] 徐佳俊，杨飏，姚天昉，等. 基于 LDA 模型的论坛热点话题识别和追踪[J]. 中文信息学报，2016，30(1)：43-50. (Xu Jiajun, Yang Yang, Yao Tianfang, et al. LDA Based Hot Topic Detection and Tracking for the Forum [J]. Journal of Chinese Information Processing, 2016, 30(1): 43-50.)

[19] 张良均，王路，谭立云，等.Python 数据分析与挖掘实战[M]. 机械工业出版社，2015. (Zhang Liangjun, Wang Lu, Tan Liyun, et al. Python Practice of Data Analysis and Mining [M]. Machinery Industry Press, 2015.)

[20] jieba [CP/OL].[2016-11-23]. http://www.oschina.net/p/jieba.
[21] 哈尔滨工业大学停用词词典[OL]. [2016-11-23]. http://more.datatang.com/data/13281. (Stop Word Dictionary by Harbin Institute of Technology [OL]. [2016-11-23]. http://more.datatang.com/data/13281.)
[22] JGibbLDA: A Java Implementation of Latent Dirichlet Allocation (LDA) Using Gibbs Sampling for Parameter Estimation and Inference [CP/OL]. [2016-11-23]. http://sourceforge.net/projects/jgibblda.

作者贡献声明:

黄孝喜, 方小飞, 谌志群: 提出研究思路, 设计研究方案;
方小飞, 黄孝喜, 王荣波: 分析数据, 进行试验, 论文起草;
王小华, 黄孝喜, 谌志群: 论文最终版本修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: 1484514227@qq.com。
[1] 方小飞. mobiledata.zip. 移动投诉文本.
[2] 方小飞. dict.txt. 投诉关键词词典.

收稿日期: 2016-11-10
收修改稿日期: 2016-12-18

Identifying Hot Topics from Mobile Complaint Texts

Fang Xiaofei [1]　Huang Xiaoxi [1]　Wang Rongbo[1]　Chen Zhiqun[1]　Wang Xiaohua [1, 2]
[1](Department of Computer Science, Hangzhou Dianzi University, Hangzhou 310018, China)
[2] (China Jiliang University, Hangzhou 310018, China)

Abstract: [**Objective**] This paper aims to extract valuable information from large amount of complaint texts with the help of Chinese message processing technologies. [**Methods**] First, we analyzed the characteristics of the complaint texts, and then clustered them by k-means algorithm. Second, we extracted topics from the texts of each category with the LDA model. In the mean time, we calculated the weight of the word of each topic, as well as the mean of document probability distribution. Third, we analyzed topics with the highest means and used the document supporting rates to identify the trending ones. [**Results**] The document supporting rates of the topics extracted by this study was three times higher than the average ones. [**Limitations**] We did not investigate the semantic relationship among the topics. [**Conclusions**] The LDA model is an effective method to detect hot topics of the mobile complaints and indicates some future studies.

Keywords: Mobile Complaints　k-means　Topic Detection　LDA Model